

A

PATENT

## IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

## Patent Application Transmittal

Assistant Commissioner for Patents  
Washington, D.C. 20231

Sir:

Transmitted herewith for filing is the Patent Application of:

Inventor: Thomas Stuermer

For: PROCESS FOR THE AUTOMATIC GENERATION OF A TEXTUAL EXPRESSION  
FROM A SEMANTIC REPRESENTATION USING A COMPUTER SYSTEM

Enclosed are:

- ☒ 13 pages of specification and 1 page of abstract  
☒ 2 sheets of drawings  
☒ A Declaration and Power of Attorney  
☐ An Information Disclosure Statement and form PTO-1449  
☒ A certified copy of a Japanese application, filed on October 29, 1998,  
 Serial No. 198 49 855.1  
☒ An assignment of the invention to International Business Machines Corporation,  
 Armonk, New York 10504 and Recordation form PTO-1595

The filing fee has been calculated as follows:

For:	No. Filed	No. Extra
Basic Fee		
Total Claims	9 -20=	0
Indep. Claims	1 -3=	0
<input type="checkbox"/> Multiple Dependent Claim Presented		

Other Than Small Entity

Rate	Fee
	\$ 760.00
x \$18.00=	\$ .00
x \$78.00=	\$ .00
\$260.00	\$ .00
<b>TOTAL</b>	<b>\$760.00</b>

## EXPRESS MAIL CERTIFICATE

Express Mail Label No.: EE205385011US

Date: June 30, 1999

I hereby certify that I am depositing the enclosed or  
attached paper with the U.S. Postal Service "Express  
Mail Post Office to Addressee" service on the above  
date, addressed to the Assistant Commissioner for  
Patents, Washington, D.C. 20231.

*Owen J. Gamon*  
Owen J. Gamon

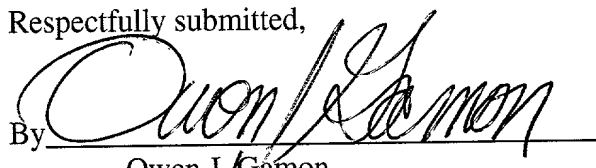
Patent Application Transmittal  
Attorney Docket No.: GE998-005

Deposit Account Authorization:

- ☒ Please charge Deposit Account No. 09-0465 in the amount of 760.00. A duplicate copy of this sheet is enclosed.
- ☒ The Commissioner is hereby authorized to charge payment of the following fees associated with this communication or credit any overpayment to Deposit Account 09-0465. A duplicate copy of this sheet is enclosed.
- ☒ Any additional filing fees required under 37 C.F.R. §1.16.
- ☒ Any patent application processing fees under 37 C.F.R. §1.17.

Respectfully submitted,

By

  
Owen J. Gamon  
Registration No. 36,143

Date: June 30, 1999

IBM Corporation  
Intellectual Property Law, Dept. 917  
3605 Highway 52 North  
Rochester, MN 55901

(507) 253-4660 voice  
(507) 253-2382 fax

**Process for the automatic generation of a textual expression from a semantic  
representation using a computer system**

5

**Field**

The invention relates to a process for the automatic generation of a textual expression from a semantic representation using a computer system.

10

**Background**

With an ever-increasing number of applications of a computer system it is advantageous or even necessary to communicate certain contents to the user. In the case of a timetable-information-system for example the computer-system should be able to show the desired timetable-information as text to the user. The text should contain not only the departure times, for example, but should provide the user with all essential information in the form of complete sentences or other textual expressions. This information can be shown on a screen in writing or can be output acoustically by way of a corresponding natural language output over a loudspeaker.

20

The production of text, especially the generation of sentences through a computer system is known.

25

Thus it is possible, that certain whole sentences are stored in a memory of the computer system, so that they can be called up again by individual programs and displayed. For example the sentence "File not found." can be stored and called up again at the appropriate time. Such procedures for the generation of sentences are also known as "canned text systems". This procedure is obviously not useable for a large number of sentences or for generating a variable sentence.

30

It is also possible, that only part of a certain sentence, which is variably completed on its retrieval from the computer system, is stored in the memory. For example, the sentence "the desired telephone number is -.." can be stored in the memory of the computer-system. On its retrieval, the computer system can add in the relevant telephone number and can output the

complete sentence to the user. Such a system for the sentence generation is also known as a “template system”. While it undoubtedly allows a certain variable sentence generation, it is nevertheless still very restricted in its flexibility.

5           Finally it is possible, with help of knowledge-based linguistic systems, to generate diverse sentences variably through a computer system. With it, so-called semantic representations which, for example, contain the information to be given to the user are determined by the computer system in a non-sentence-bound and non-grammatical format. This must be changed by the computer system into textual expressions or complete sentences. 10 Such systems for the generation of textual expressions from semantic representations necessitate a substantial expenditure in storage area, computing time and development activity, so that they have so far only been available for certain, narrowly restricted generation tasks.

### Summary

15           It is the object of the invention to provide a procedure for the automatic generation of a textual expression from a semantic representation through a computer system, which can be employed for as large a number of semantic representations as possible and as flexibly as possible, which nevertheless necessitates only a justifiable expenditure in memory and computing time.

20           This object is achieved by the invention by a procedure for the automatic generation of an expression from a semantic representation in accordance with Claim 1.

25           The invention uses a statistical model for the investigation of an expression from a semantic representation. The determined expression is here the most likely expression, which can be assigned to a pre-determined semantic representation on the basis of the statistical model.

30           This allows the advantages of the known “canned text systems”, especially their fast accessibility, to be largely retained. Simultaneously, however, the number of expressions available becomes virtually unlimited.

The development of program code for the generating steps is widely separated from the statistical model and its development. In this way, the expansion of the procedure and its adaptation where necessary is essentially facilitated.

5           A change, for example, from the German to the English language can be achieved by an exchange of the German language by an English language translation within the statistical model alone, which means a considerable reduction in development costs.

10           For the determination of the statistical model, known, existing procedures or programs can be used, which also leads to a reduced expenditure in the development of the whole procedure.

15           Also it is possible that, for semantic representations not contained in the statistical model, at least useable textual expressions are generated. Consequently, the procedure is also suitable for the processing of new semantic representations.

20           In addition, the procedure in accordance with the invention exhibits no especially high storage area requirement and is also not critical in respect of its computing time for generation. Consequently, it does not possess the disadvantages of knowledge-based linguistic systems.

25           Through the conversion of the semantic representations and the associated expressions into an internal format, it is possible to simplify the corpus and the statistical model deriving from it. In particular, it further decreases the amount of memory required. Also irrelevant information can be filtered out for the statistical processing. The model is thus improved and computing is accelerated.

30           Through the introduction of generalizations, it is possible to widen the number of the expressions available practically indefinitely, without thereby significantly increasing the demands on memory and computing time. Conversely, a corpus becomes more compact by the summarizing of equivalent entries and the statistical model to be calculated more robust and more accurate.

**Brief Description of the Drawings**

Further characteristics, uses and advantages of the invention result from the following description of examples of embodiments of the invention, which are shown in figures of the drawing. Here, all features described or represented alone or in any combination, form the subject matter of the invention, independent of their summary in the patent claims or their reference back and independent of their formulation or representation in the description or in the drawing.

Figure 1 shows a diagrammatic representation of a method for the investigation of a statistical model.

Figure 2 shows a diagrammatic representation of a method for the generation of expressions on the basis of the statistical model of figure 1.

Figure 3 shows a diagrammatic representation of a method for the generalization of a corpus, from which a generalized statistical model in accordance with figure 1 is determined.

Figure 4 shows a diagrammatic representation of a method for the generation of expressions on the basis of the generalised statistical model of figure 3.

**Detailed Description**

The starting point for the procedure for the calculation of a statistical model in accordance with figure 1, and also for the generation of expressions in accordance with figure 2, are pairs of semantic representations and associated expressions. From a multiplicity of such pairs a statistical model is determined in a preparatory step - the procedure in accordance with figure 1 -, which subsequently - in the procedure in accordance with figure 2 - is searched again for the purpose of generating sentences.

The manner of representing semantic representations and associated expressions is pre-determined and is marked as external. This external method of showing the pairs of semantic representations and associated expressions can, for example, be shown as follows:

Semantic representation, external mode of representation:

from\_until(15,14),

```

def(17,i2,l1,htl90),
mofy(13,il,apr),
dofm(14, i2 ,i1,27),
dofm(15,i2,i3,23)...

```

5 associated expression, external mode of representation:

“from the 23rd to the 27th of April”

10 It is possible, that this external mode of representation is produced independently by other programs of the computer-system. Also it is possible, that this or any other mode of representation results from the application of other programs of the computer-system. For example the mode of representation can itself be given by the application of a translation program or a speech-recognition program or the like which already uses the corresponding external mode of representation.

15 In the external mode of representation a multiplicity of pairs of semantic representations and of associated expressions are entered into the computer system and there stored in memory.

20 The pairs of semantic representations and associated expressions can either be input into the computer-system manually. In this case, a person can input into the computer system just those semantic representations and associated expressions, which are meaningful or necessary for the relevant application. It is also possible, that the pairs of semantic representations and associated expressions from another program, for example from the translation program or the speech-recognition program are converted and stored in the computer-system.

25 All pairs of semantic representations and associated expressions input into the computer system represent a corpus. The more extensive this corpus is, the more extensive also are the possibilities of a later generation of expressions through the computer system.

30 Starting with this corpus the computer system determines a statistical model in accordance with the procedure of figure 1.

For this, the corpus is first converted in a step 11 from the external mode of representation into an internal format. The pairs  $C$  are thus represented in pairs of  $C_{\text{internal}}$ . To carry out this step, a negative list 12 and a translation table 13 are used. The internal format represents a simplification of the external mode of representation of the corpus, which is directed towards the determination of a statistical model to be carried out or is restricted to this.

The negative list describes information, which may be contained in the external representation, but which is irrelevant or even damaging for the statistical model. This information is not found in the internal representation. The translation table produces the reference from parts of the external representation to parts of the internal representation.

For example, the initially mentioned pair of a semantic representation and the associated expression are converted from the external mode of representation into the internal format as follows:

Semantic representation, internal format:

from\_until,def,i1,i2,i3,11,...,17,  
dofm(27th,23rd),mofy(April)

associated expression, internal format:

from,the,23rd,to,the,27th,of,April

The corpus only existing in the internal format forms the basis for the determination of a statistical model. For this in a step 14 a number of questions  $Q_{\text{start}}$  were determined by the computer-system, with which the single entries of the corpus can be classified. In addition it is possible, that a number of initialising questions 15 can also be pre-determined manually.

From these questions, the computer system in a step 16 calculates on the basis of known procedures a statistical model to the existing corpus. With this, questions are generated for each component of the existing semantic representations of the computer-system, so that



the relevant components are classified. This way the computer system produces a statistical decision-tree model T, for example.

Regarding the production of these questions, pre-determined boundary conditions can play a role, for example, that the decision-tree-model produced should preferably be properly balanced or the like.

The model T calculated in the step 16 is stored in the computer system. The investigation of this model T, and hence the procedure in accordance with figure 1, is gone through once, in fact before the actual generation of expressions in accordance with the procedure of figure 2. The last-named procedure then takes place on the basis of the calculated model T.

In the procedure in accordance with figure 1 - pairs C of semantic representations and their associated expressions are therefore converted from the external mode of representation into the internal format, in order that the statistical model T can then be determined from the pairs C<sub>internal</sub>. These steps are carried out automatically by the computer system.

On the basis of the statistical model T the computer system can generate expressions in accordance with the procedure of figure 2.

For this, a semantic representation is supplied to the computer system as input i. This is converted in a step 21 first again from the external mode of representation into the internal format. For this the negative list 12 and the translation table 13 are again used. After the step 21 the semantic representation is consequently available as input i<sub>internal</sub>.

In a step 22 the statistical model T is searched. This is, for example, - as already explained - stored in the computer system as decision-tree model 23. The above-mentioned questions are consequently put to the semantic representation converted into the internal format and will be followed, for example, in parallel, by a certain number of alternative answers, which possess the highest likelihood on the basis of model T and input i.

In this way a statistical search in the decision-tree model is carried out by the computer system, by which the most suitable expression for the pre-determined semantic representation

is determined. This expression represents an output o<sub>internal</sub>, which is still in the internal format.

In a step 24 the internal format of the output o<sub>internal</sub> is again re-converted to the external representation mode, so that the expression produced is henceforth available as output o in the external representation mode.

Thus, the input i is converted to the internal format with the procedure in accordance with figure 2. Then the input i<sub>internal</sub> is classified on the basis of the statistical model, in order to determine the most suitable output o<sub>internal</sub>. This is finally re-converted to the external representation in output o. These steps are carried out automatically by the computer system.

It is possible, if necessary, after the determination of the most suitable expression, for the correctness of this expression to be checked again by the computer system with help of additional procedures.

The procedure in accordance with figure 2 and if necessary the named additional procedures is gone through for each input i, so that for each input i, and hence for each semantic representation, an associated output O, and hence an associated expression, is generated by the computer-system.

To increase the capability of the above-described procedures it is possible to generalise the corpus consisting of pairs of semantic representations and associated expressions.

For this purpose, certain parts of associated semantic representations and the associated expressions, which can be generalised and which appear repeatedly, are generalised by the use of so-called patterns and templates. So, for example, a specific month is no longer adopted into the corpus as such but as a generalised month. In the generation of expressions, the assignment of the generalised month is stored to the specific month in the computer system beforehand and then subsequently re-introduced into the expression produced.

By way of example, a generalisation of the semantic representations and the associated expressions can be carried out as follows:

Semantic representations:

associated expressions:

external representation mode:

mofy(13,i1,apr),  
dofm(14,i2,i1,27),  
dofm(16,i2,i3,23),

“From the 23rd to  
the 27th of April”

internal format:

from\_until,def,i1,i2,i3,  
11,...,17,  
dofm(27th,23rd),  
mofy(April)

from,the,23rd,to,  
the,27th,of,April

generalised format:

from\_until,def, i1,i2,i3,  
11,...,17,  
dofm(Vdofm1,Vdofm2)  
mofy(Vmofy1)

from,the,Vdofm2,to  
the,vdofm2,of,Vmofy1

In the generalised format, the specific month “April”, for example, is replaced by the generalised month “Vmofy1”. The corpus therefore no longer contains the specific month “April” but only the generalised month “Vmofy1”. The same applies to the specific month “June”, for example, which is also represented in the corpus by the same generalised month “Vmofy1”. Through the generalisation, the specific months of the year “January, February,..., December” are thus in this case replaced by the generalised month, “Vmofy1”.

Starting from the pairs Cinternal of semantic representations and associated expressions present in the internal format, the computer-system then determines in accordance with figure 3 pairs Cgeneralised, that are available in a generalised format. This generalisation is carried out for all individual pairs cinternal, which in each case are converted to pairs cgeneralised.

Each pair of the corpus is checked by the computer system for possible substitutions using the procedure in accordance with figure 3. The substitutions can be supplied manually or may also be determined automatically by the computer system if necessary. The substitutions are stored in the memory of the computer system. In figure 3 the substitutions are marked with the reference number 31.

On the basis of these substitutions 31 the semantic representation *i*internal of a specific pair *c*internal present in the internal format is converted in a step 32 into a semantic representation present in the generalised format *i*pattern. In a corresponding manner, the associated expression *o*internal present in the internal format of the specific pair *c*internal is converted in a step 33 into an expression *o*template present in the generalised format. In general, for each pair *c*internal present in the internal format a pair *c*generalised present in the generalised format is produced, consisting of the respective semantic representation *i*pattern and the associated expression *o*template, each in the generalised format.

Subsequently, a generalised statistical model *T'* is produced from the pairs *C*generalised by the computer system in the same manner, as has already been explained in connection with the procedure in accordance with figure 1. The decision-tree model produced differs from that of figure 1 mainly in that it contains the stored substitutions 31, which thus contains their substitute "Vmofy1" instead of the specific months "April" or "June" or the like.

On the basis of the generalised statistical model *T'* the computer-system can generate expressions in accordance with the procedure of figure 4.

To do this, a semantic representation is supplied to the computer system as input *i*. This is first converted again in a step 41 from the external representation mode into the internal format. For this, the negative list 12 and the translation table 13 are again used. After the step 41, the semantic representation is thus available as input *i*internal.

In a step 42, the generalised statistical model *T'* is searched. This is stored in the computer-system as the decision-tree model 43. This way a statistical search in the decision-tree model is carried out by the computer system, which determines the most suitable

expression for the pre-determined semantic representation. This expression may - because of the generalisations underlying the statistical model T - contain substitutions.

5 In a step 44 the computer system, at approximately the same time as step 42, stores in its memory the substitutions actually occurring in the semantic representations internal with their respective specific content.

10 The temporary expressions produced in step 43 and the substitutions stored in step 44 are revised in a step 45 in such a way that the existing substitutions are again replaced. Thus the stored content is now put into the temporary expression again instead of the substitution, so that the final expression corresponds to the initial semantic representation. This expression represents an output ofilled, that is still in the internal format.

15 In a step 46 the internal format of the output ofilled is re-converted into the external representation mode again, so that the expression produced is now available as output o in the external representation mode.

What is claimed is:

# CLAIMS

- 1 1. A method for the automatic generation of a textual expression from a semantic  
2 representation, comprising the computer-executed steps of:  
  
3 building a statistical model from a plurality of pre-determined pairs of semantic  
4 representations and associated expressions; and  
  
5 producing a first associated expression from a first semantic representation using the  
6 statistical model.
- 1 2. The method of claim 1, wherein the building step further comprises:  
2  
3 converting the pairs of semantic representations and associated expressions from an  
4 external format into an internal format.
- 1 3. The method of claim 2, wherein the converting step further uses a negative list and a  
2 translation table.
- 1 4. The method of claim 3, wherein the negative list describes information in the external  
2 format that is irrelevant to the statistical model.
- 1 5. The method of claim 3, wherein the negative list describes information in the external  
2 format that is damaging to the statistical model.
- 1 6. The method of claim 3, wherein the converting step further comprises:  
2 determining which information in the external format is present in the negative list and  
3 which information in the external format is absent in the negative list;  
4 converting information that is absent in the negative list from the external format to the  
5 internal format using the translation table; and  
6 refraining from converting information that is present in the negative list.

1 7. The method of claim 2, wherein the building step further comprises:  
2 determining a plurality of questions;  
3 classifying the information in the internal format using the plurality of questions; and  
4 calculating the statistical model from the internal format using the plurality of  
5 questions.

1 8. The method of claim 7, wherein the determining step further comprises:  
2 determining the plurality of questions from pre-determined boundary conditions.

1 9. The method of claim 1, further comprising:  
2 storing the statistical model as a decision-tree model.

**Process for the automatic generation of a textual expression from a semantic representation using a computer system**

5

**ABSTRACT**

A procedure for the automatic generation of a textual expression from a semantic representation by a computer-system is described. With the procedure, a statistical model is determined by the computer-system on a plurality of pre-determined pairs of semantic representations and associated expressions and stored. A semantic representation, from which an associated expression is determined by the computer system by means of the statistical model, is presented to the computer system. These steps are repeated by the computer system for further semantic representations if necessary.



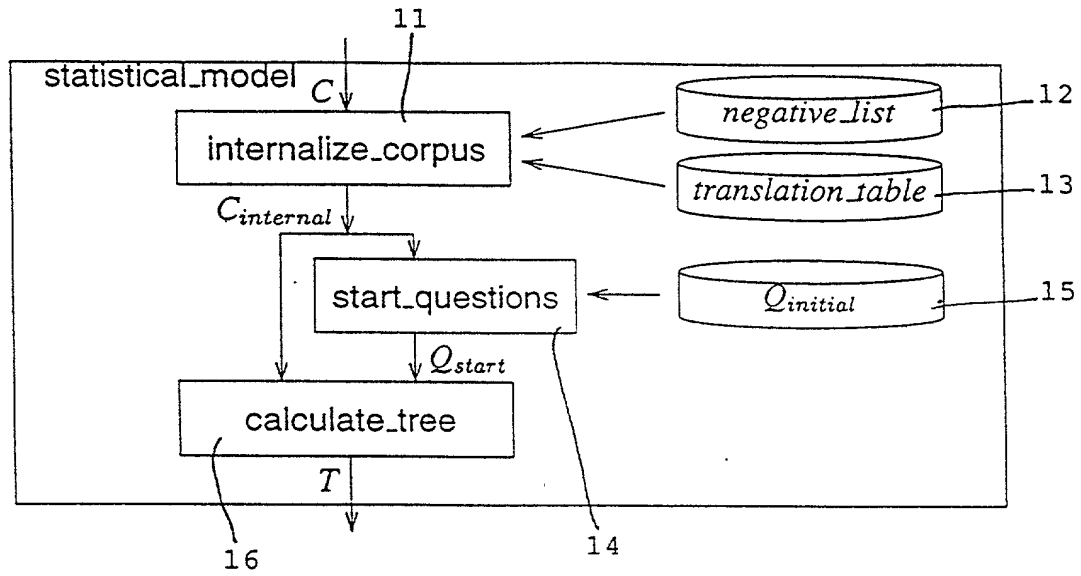


FIG. 1

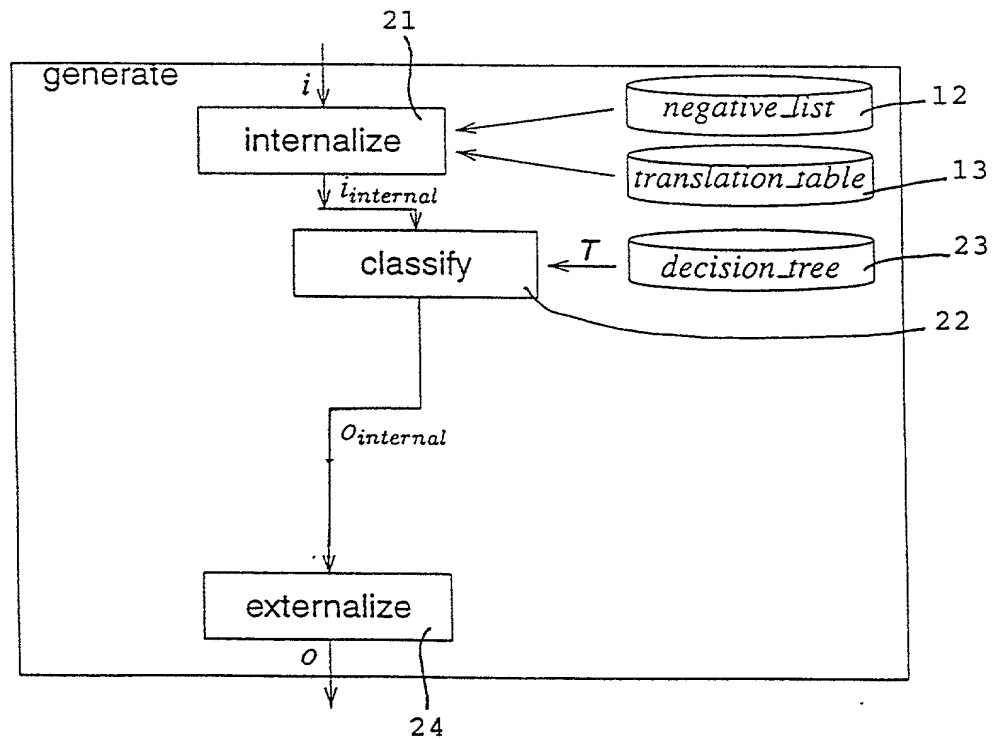


FIG. 2

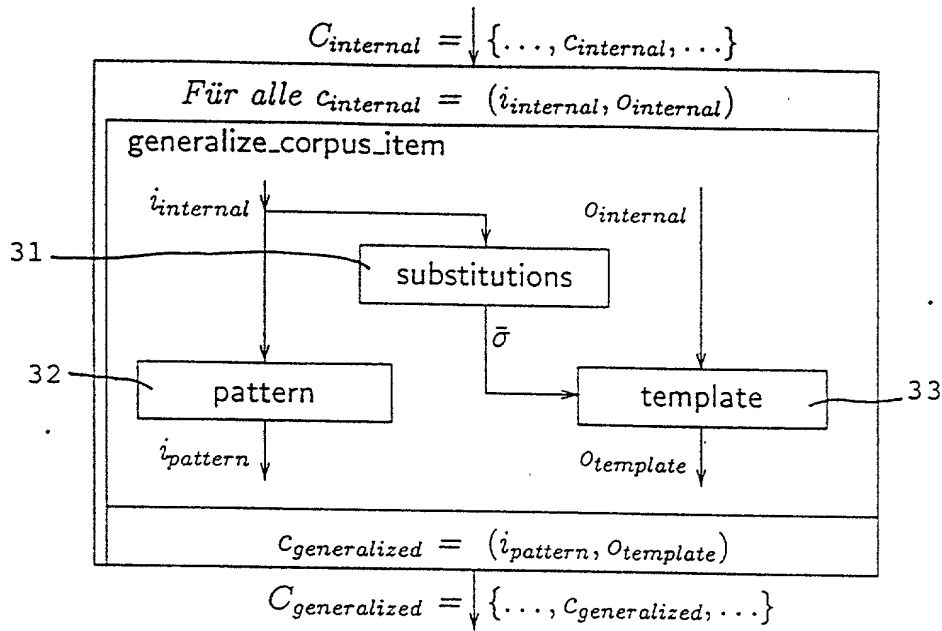


FIG. 3

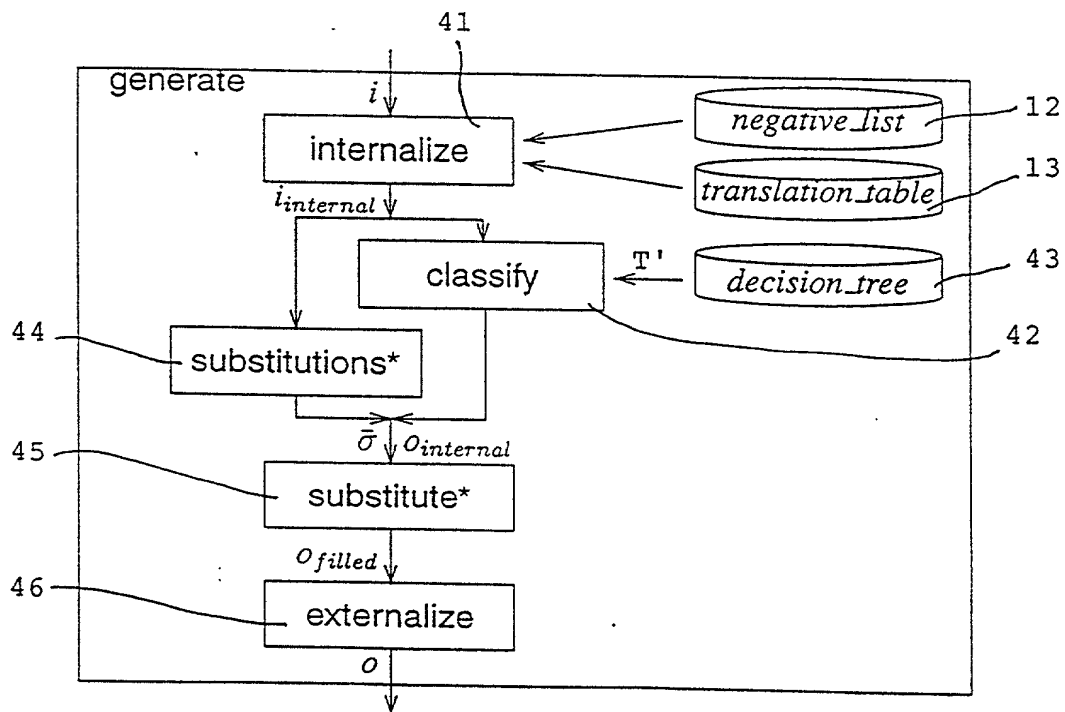


FIG. 4

---

**DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION**Docket No.: GE998-005

---

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name. I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

**PROCESS FOR THE AUTOMATIC GENERATION OF A TEXTUAL EXPRESSION  
FROM A SEMANTIC REPRESENTATION USING A COMPUTER SYSTEM**

the specification of which (check one)

X  is attached hereto.

\_\_\_\_\_ was filed on \_\_\_\_\_ as

Application Serial No. \_\_\_\_\_ and

was amended on \_\_\_\_\_

(if applicable)

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the patentability of this application in accordance with Title 37, Code of Federal Regulations, §1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, §119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

Prior Foreign Application(s)

Priority Claimed

198 49 855.1    Germany    29/10/1998     X  YES    NO  
(Number)                      (Country)                      (Day/Month/Year Filed)

I hereby claim the benefit under Title 35, United States Code, §120 of any United States Application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, §112, I acknowledge the duty to disclose information material to the patentability of this application as defined in Title 37, Code of Federal Regulations, §1.56(a) which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

\_\_\_\_\_(NONE)\_\_\_\_\_  
(Application Serial No.) (Filing Date) (Status) (Patented, Pending, Abandoned)

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

POWER OF ATTORNEY: As a named inventor I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith.  
(List name and registration number)

Matthew J. Bussan - 33,614	Roy W. Truelson - 34,265
Owen J. Gamon - 36,143	John E. Hoel - 26,279
Pryor A. Garnett - 32,136	Christopher A. Hughes - 26,914
Steven W. Roth - 34,712	Edward A. Pennington - 32,588
	Joseph C. Redmond, Jr. - 18,753

Send Correspondence to: Owen J. Gamon  
IBM Corporation, Dept. 917  
3605 Highway 52 North  
Rochester, MN 55901-7829

Direct Telephone Calls to: Owen J. Gamon  
Area Code 507-253-4660

Full name of sole or first Inventor

Thomas Stuermer  
Inventor's signature

Date

June 09, 1999

Residence

Habichtsweg 24, D-66123 Saarbruecken, Federal Republic of Germany

Citizenship

German

Post Office Address

Same as above

☒ This declaration ends with this page.